# *Appendix B - Statistical Methods used in the Idaho National Engineering and Environmental Laboratory Annual Site Environmental Report*

*M. Case - S. M. Stoller Corporation*
*J. Einerson - Bechtel/BWXT Idaho, LLC.*

Relatively simple statistical procedures are used to analyze the data collected by the Idaho National Engineering and Environmental Laboratory (INEEL) Environmental Surveillance, Education and Research (ESER) program. This appendix presents the guidelines used to evaluate sample results.

## *Guidelines for Reporting Results*

The results reported in the quarterly and annual reports are assessed in terms of data quality and statistical significance with respect to laboratory analytical uncertainties, sample locations, reported INEEL releases, meteorological data, and worldwide events that might conceivably have an effect on the INEEL environment.

## *Initial Screening*

First, field collection and laboratory information are reviewed to determine identifiable errors that would invalidate or limit use of the data. Examples of field observations which could invalidate the result include insufficient sample volume, torn filters, or mechanical malfunction of sampling equipment.

The analytical laboratory also qualifies the results and may reject them for reasons such as:

- the uncertainty is too high to be accepted by the analyst;
- the radionuclide has no supporting photopeaks to make a judgment;
- the photopeak width is unacceptable by the analyst;
- the result is below the decision critical level;
- other radionuclides display gamma-ray interferences;
- a graphical display of analyzed photopeaks showed unacceptable fitting results;
- there is no parent activity, therefore the state of equilibrium is unknown and the radionuclide could not be quantified; and
- the radionuclide is a naturally-occurring one with expected activity.

Evidence of laboratory cross-contamination or quality control issues could also disqualify a result (see Chapter 10.)

Data that pass initial screening are further evaluated prior to reporting.

### *Reporting Levels*

It is the goal of the ESER program to minimize the error of saying something is not present when it actually is, to the extent that is reasonable and practicable. This is accomplished through the use of the uncertainty term, which is reported by the analytical laboratory with the sample result. For radiological data, individual analytical results are usually presented in this report with plus or minus one sample standard deviation (± 1s). The sample standard deviation is obtained by propagating sources of analytical uncertainty in laboratory measurements. The uncertainty term, "s," is an estimate of the population standard deviation "σ," assuming a Guassian or normal distribution. The approach used by the ESER program to interpret individual analytical results is based on guidelines outlined by the U.S. Geological Survey (USGS) in Bartholomay et al. (2000), which are based on methodology proposed by Currie (1984). Most of the following discussion is from Bartholomay et al. (2000).

Laboratory measurements are made on a target sample and on a laboratory-prepared blank. Instrument signals for the sample and blank vary randomly about the true signals. Two key concepts characterize the theory of detection: the "critical value" (or "critical level" or "criterion of detection") and the "minimum detectable value" (or "detection limit" or "limit of detection"). The critical level and minimum detectable concentration are based on counting statistics alone

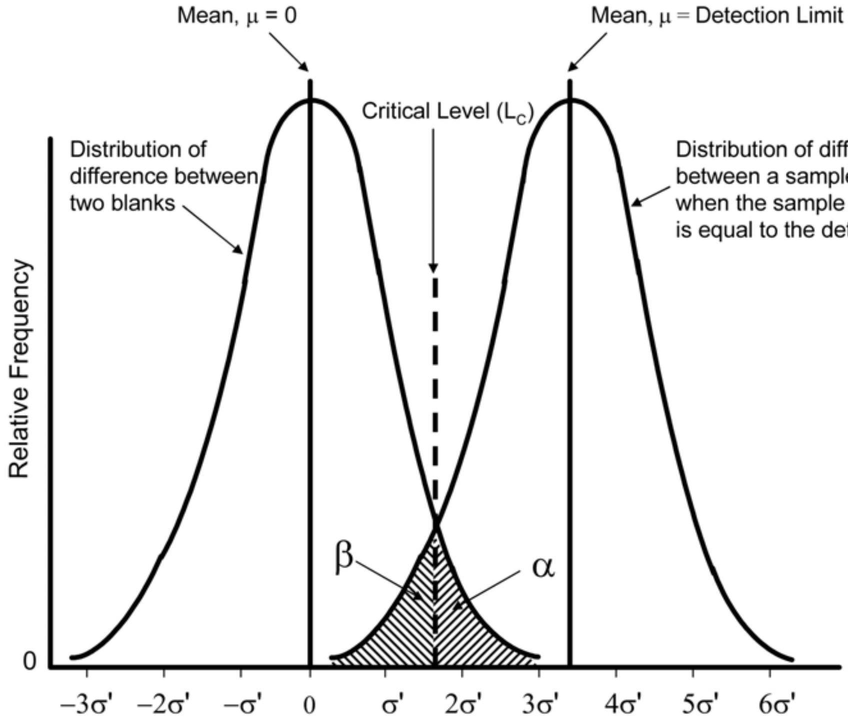
Mean, μ = 0
Mean, μ = Detection Limit ($L_D$)
Critical Level ($L_C$)
Distribution of difference between two blanks
Distribution of difference between a sample and a blank when the sample concentration is equal to the detection limit
Relative Frequency
β
α
0
−3σ'
−2σ'
−σ'
0
σ'
2σ'
3σ'
4σ'
5σ'
6σ'

**Figure B-1. Illustration of the relation of the criterion of detection (critical level) and the limit of detection (detection limit). Errors of the first kind (false negatives) are represented by the value of α, whereas errors of the second kind (false positives) are represented by the value of β. (from Currie 1984)**

and do not include systematic or random errors inherent in laboratory procedures. Figure B-1 illustrates these terms.

The critical level ($L_C$) is the minimum significant value of an instrument signal or concentration that can be discriminated from the signal or concentration observed for the blank such that the decision can be made that the radionuclide was detected. The decision "detected" or "not detected" is made by comparison of the estimated quantity $(\hat{L})$ with $L_C$. A result falling below $L_C$ triggers the decision "not detected". That is, the probability distribution of possible outcomes, when the true net signal is zero, intersects $L_C$ such that the fraction 1-$\alpha$, where $\alpha$ is the error of the first kind (false positive), corresponds to the correct decision "not detected". Typically $\alpha$, is set equal to 0.05. Using algorithms in Currie (1984) that are appropriate for our data, the $L_C$ is 1.65s or approximately 2s. At this level, there is about a 95 percent probability that the correct decision—not detected—will be made. Given a large number of samples, as many as 5 percent of the samples with measured concentrations larger than or equal to 2s, which were concluded as being detected, might not contain the radionuclide (i.e., a false positive).

Once the critical level has been defined, the minimum detectable concentration (MDC), or detection level ($L_D$), may be determined. Using the equations in Curries (1984), concentrations that equal 3.29s, or approximately 3s, represent a measurement at the minimum detectable concentration. For true concentrations of 3s or larger, there is 95 percent or larger probability that the radionuclide was detected in a sample. In a large number of samples, the conclusion, not detected, will be made in 5 percent of the samples that contain true concentrations at the minimum detectable concentration of 3s. These are referred to as false negatives or errors of the second kind.

True radionuclides concentrations between 2s and 3s have larger errors of the second kind. That is, there is a larger-than-five-percent probability of false negative results for samples with true concentrations between 2s and 3s. Although the radionuclide might have been detected, such detection may not be considered reliable; at 2s, the probability of a false negative is about 50 percent.

In this report, radionuclide concentrations less than 3s are considered to be below a "reporting level." Concentrations equal to or above 3s are considered to be detected with confidence. Results between 2s and 3s are considered to be "questionable" detections. Results less than or equal to 2s are reported as “undetected.” Each result is reported with the associated 1s uncertainty value for consistency with other INEEL reports.

## *Statistical Tests used to Assess Data*

An example set of data are presented here to illustrate the statistical tests used to assess data collected by the ESER contractor. The dataset used are the gross beta environmental surveillance data collected from January 8, 1997, through December 26, 2001. The data were collected weekly from several air monitoring stations located around the perimeter of the INEEL and air monitoring stations throughout the Snake River Plain. The perimeter locations are termed "boundary" and the Plain locations are termed "distant." There are seven boundary locations: Arco, Atomic City, Birch Creek, FAA Tower, Howe, Monteview, and Mud Lake, and five distant locations: Blackfoot, Blackfoot Community Monitoring Station (CMS), Craters of the Moon,

Idaho Falls, and Rexburg CMS. The gross beta data are of the magnitude $10^{-15}$. To simplify the calculations and interpretation, these have been coded by multiplying each measurement by $10^{15}$.

Only portions of the complete gross beta data set will be used. The purpose of this task is to evaluate and illustrate the various statistical procedures, and not a complete analysis of the data.

## *Test of Normality*

The first step in any analysis of data is to test for normality. Many standard statistical tests of significance require that the data be normally distributed. The most widely used test of normality is the Shapiro-Wilk W test (Shapiro, S.S. and M.B. Wilk 1965). The Shapiro-Wilk W test is the preferred test of normality because of its good power properties as compared to a wide range of alternative tests (Shapiro, S.S. et al. 1968). If the W statistic is significant ($p<0.00001$), then the hypothesis that the respective distribution is normal should be rejected.

Graphical depictions of the data should be a part of any evaluation of normality. The following histogram (Figure B-2) presents such a graphical look along with the results of the Shapiro-Wilk W test. The data used for the illustration are the five years of weekly gross beta measurements for the Arco boundary location. The W statistic is highly significant ($p<0.0001$) indicating that the data are not normally distributed. The histogram shows that the data are asymmetrical with right skewness. This suggests that the data may be lognormally distributed. The Shapiro-Wilk W test can be used to test this distribution by taking the natural logarithms of each measurement and calculating the W statistic. Figure B-3 presents this test of lognormality. The W statistic is not significant ($p=0.80235$) indicating that the data are lognormal.

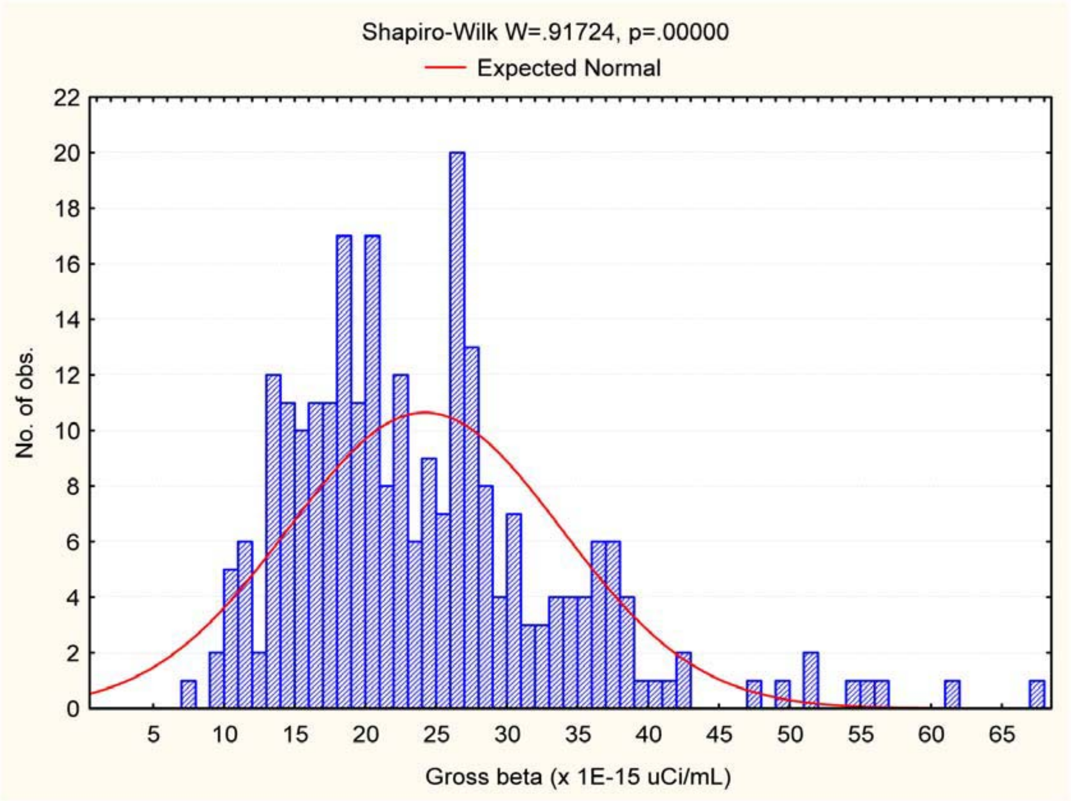

**Figure B-2. Test of normality for Arco gross beta data.**

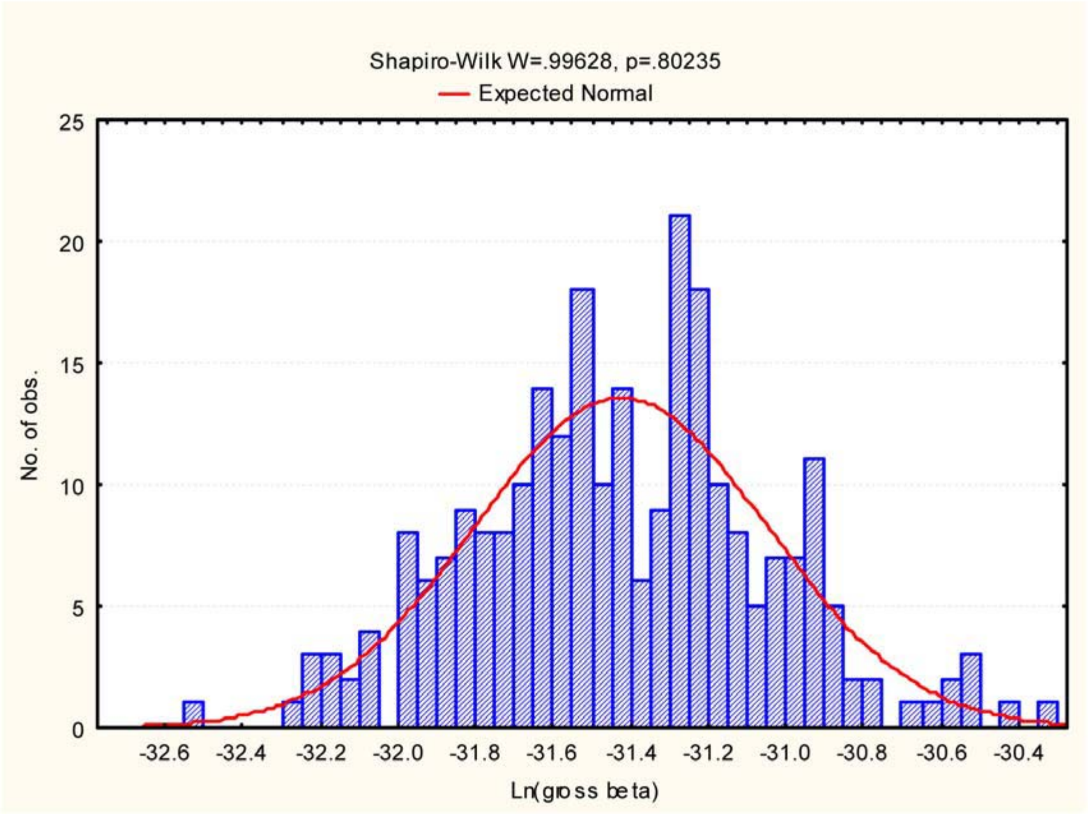

**Figure B-3. Test of log normality for Arco gross beta.**

To perform parametric tests of significance such as Student's T Test or One-Way Analysis of Variance (ANOVA), it is required that all data be normally (or lognormally) distributed. Therefore, if one desires to compare gross beta results of each boundary location, tests of normality must be performed before such comparisons are made. Table B-1 presents the results of the Shapiro-Wilk W Test for each of the seven boundary locations.

From Table B-1, none of the locations consist of data that are normally distributed and only some of the data sets are lognormally distributed. This is a typical result and a common problem when one desires to use a parametric test of significance. When many comparisons are to be made, attractive alternatives are nonparametric tests of significance.

## *Comparison of Two Groups*

For comparison of two groups, the Mann-Whitney U Test (Hollander, M. and D.A. Wolfe 1973) is a powerful nonparametric alternative to the Student's T Test. In fact, the U Test is the most powerful (or sensitive) nonparametric alternative to the T Test for independent samples; in some instances it may offer even greater power to reject the null hypothesis than the T Test. The interpretation of the Mann-Whitney U Test is essentially identical to the interpretation of the Student's T Test for independent samples, except that the U Test is computed based on rank sums rather than means. Because of this fact, outliers do not present the serious problem that they do when using parametric tests.

**Table B-1. Tests of normality for boundary locations.**

| Location | Normal | | Lognormal | |
|---|---|---|---|---|
| | W statistic | p-value | W statistic | p-value |
| Arco | 0.9172 | <0.0001 | 0.9963 | 0.8024 |
| Atomic City | 0.9174 | <0.0001 | 0.9411 | <0.0001 |
| Birch Creek | 0.8086 | <0.0001 | 0.9882 | 0.0530 |
| FAA Tower | 0.9119 | <0.0001 | 0.9915 | 0.1397 |
| Howe | 0.8702 | <0.0001 | 0.9842 | 0.0056 |
| Monteview | 0.9118 | <0.0001 | 0.9142 | <0.0001 |
| Mud Lake | 0.6130 | <0.0001 | 0.9704 | <0.0001 |

Suppose we wish to compare all boundary locations to all distant locations. Figure B-4 presents the box plots for the two groups. The median is the measure of central tendency most commonly used when there is no assumed distribution. It is the middle value when the data are ranked from smallest to largest. The 25th and 75th percentiles are the values such that 75 percent of the measurements in the data set are greater than the 25th percentile and 75 percent of the measurements are less than the 75th percentile. The large distance between the medians and the maximums seen in Figure B-4 indicate the presence of outliers. It is apparent that the medians are

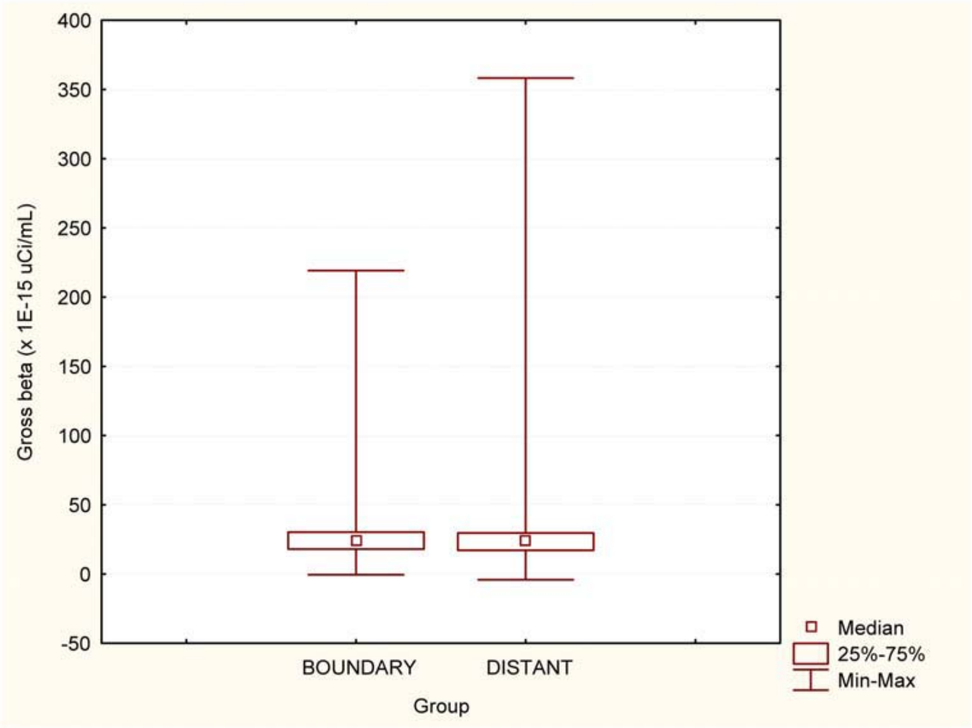

**Figure B-4. Box plot of gross beta data from boundary and distant locations.**

of the same magnitude indicating graphically that there is probably not a significant difference between the two groups.

The Mann-Whitney U test compares the rank sums between the two groups. In other words, for both groups combined, it ranks the observations from smallest to largest. Then it calculates the sum of the ranks for each group and compares these rank sums. A significant p-value ($p<0.05$) indicates a significant difference between the two groups. The p-value for the comparison of boundary and distant locations is not significant ($p=0.0599$). Therefore, the conclusion is that there is not strong enough evidence to say that a significant difference exists between boundary and distant locations.

## *Comparison of Many Groups*

Now suppose we wish to compare the boundary locations amongst themselves. In the parametric realm, this is done with a One-Way ANOVA. A nonparametric alternative to the One-Way ANOVA is the Kruskal-Wallis ANOVA (Hollander, M. and D.A. Wolfe 1973). The test assesses the hypothesis that the different samples in the comparison were drawn from the same distribution or from distributions with the same median. Thus, the interpretation of the Kruskal-Wallis ANOVA is basically identical to that of the parametric One-Way ANOVA, except that it is based on ranks rather than means.

Figure B-5 presents the box plot for the boundary locations. The Kruskal-Wallis ANOVA test statistic is highly significant ($p<0.0001$) indicating a significant difference amongst the seven boundary locations. Table B-2 gives the number of samples, medians, minimums, and maximums for each boundary location. The Kruskal-Wallis ANOVA only indicates that significant differences exist between the seven locations and not the individual occurrences of differences. If desired, the next step is to identify pairs of locations of interest and test those for significant differences using the Mann-Whitney U test. It is cautioned that all possible pairs should not be tested, only those of interest. As the number of pairs increases, the probability of a false conclusion also increases.

Suppose a comparison between Arco and Atomic City is of special interest due to their close proximity to each other. A test of significance using the Mann-Whitney U test results in a p-value of 0.7288 indicating that a significant difference does not exist between gross beta results at Arco and Atomic City. Other pairs can similarly be tested, but with the caution given above.

## *Tests for Trends over Time*

Regression analysis is used to test whether or not there is a significant positive or negative trend in gross beta concentrations over time. To illustrate the technique, the regression analysis is performed for the boundary locations as one group and the distant locations as another group. The tests of normality performed earlier indicated that the data were closer to lognormal than normal. For that reason, the natural logarithms of the original data are used in the regression analysis. Regression analysis assumes that the probability distributions of the dependent variable (gross beta) have the same variance regardless of the level of the independent variable (collection date). The natural logarithmic transformation helps in satisfying this assumption.

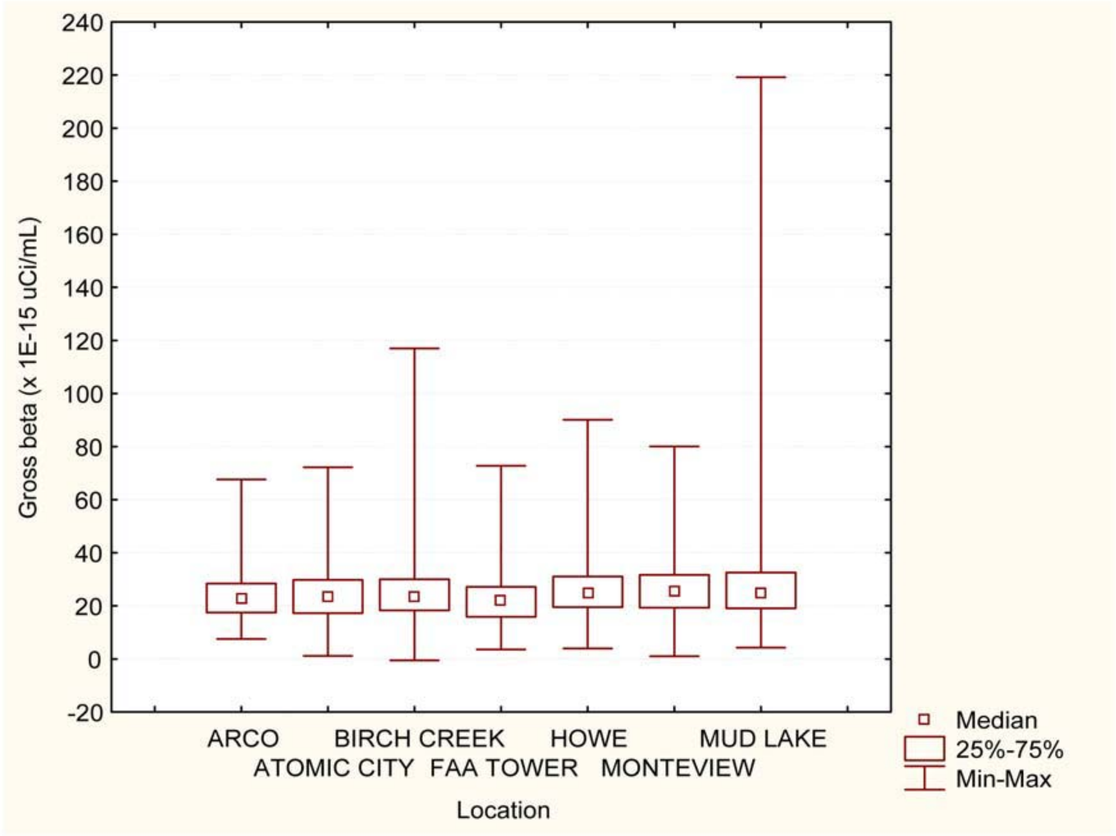

**Figure B-5. Box plot of gross beta data for each boundary location.**

**Table B-2. Summary statistics for boundary locations.**

| Location | Number of Samples | Median | Minimum | Maximum |
|---|---|---|---|---|
| Arco | 258 | 22.49 | 7.53 | 67.66 |
| Atomic City | 260 | 23.61 | 1.13 | 72.20 |
| Birch Creek | 234 | 23.15 | -0.52 | 117.00 |
| FAA Tower | 260 | 21.90 | 3.59 | 72.78 |
| Howe | 260 | 24.55 | 3.95 | 90.10 |
| Monteview | 260 | 25.30 | 1.03 | 80.10 |
| Mud Lake | 260 | 24.85 | 4.30 | 219.19 |

a. All values are × $10^{-15}$ microcuries per milliliter (µCi/mL).

Figure B-6 presents a scatterplot of the boundary data with the fitted regression line superimposed. Figure B-7 presents the same for the distant data. Table B-3 gives the regression equation and associated statistics. There appears to be slightly increasing trends in gross beta over time for both the boundary and distant locations. A look at the regression equations and correlation coefficients in Table B-3 confirm this. Notice that the slope parameter of the regression equation and the correlation coefficient are equal. This is true for any linear regression fit. So, a test of significant correlation is also a test of significant trend. The p-value associated with testing whether or not the correlation coefficient is different from zero is the same as for testing if the slope of the regression line is different from zero. For both the boundary and distant locations, the slope is significantly different from zero and positive indicating an increasing trend in gross beta over time.

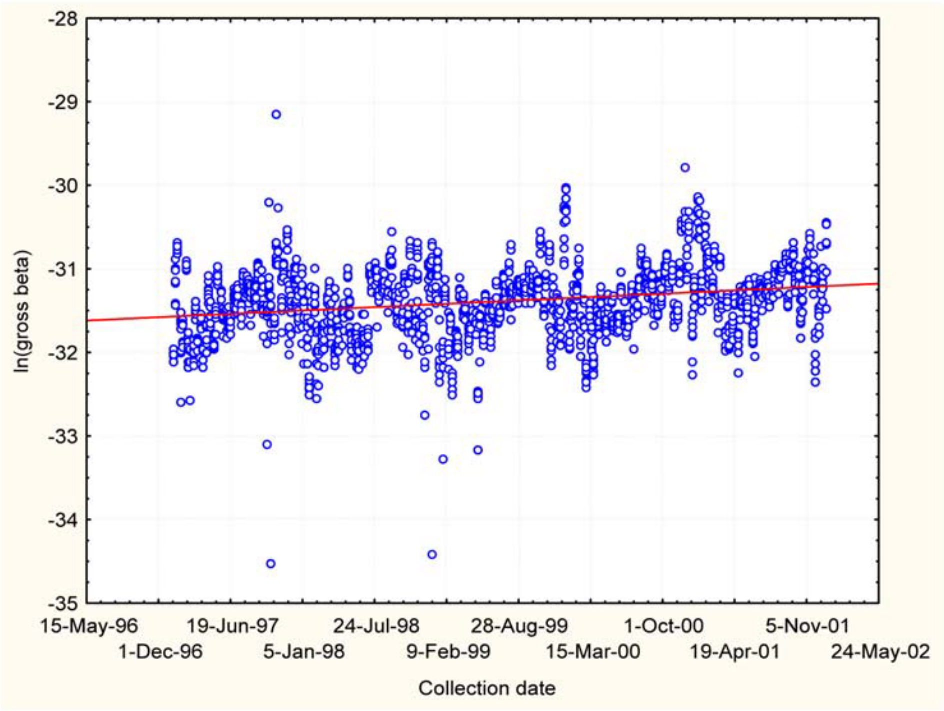

**Figure B-6. Scatter plot and regression line for ln(gross beta) from boundary locations.**

Another important point of note in Figures B-6 and B-7 is the obvious existence of a cyclical trend in gross beta. It appears as if the gross beta measurements are highest in the summer months and lowest in the winter months. Since the regression analysis performed above is over several years, we are still able to detect a positive trend over time even though it is confounded somewhat by the existence of a cyclical trend. This is important because a linear regression analysis performed over a shorter time period may erroneously conclude a significant trend, when in fact, it is just a portion of the cyclical trend.

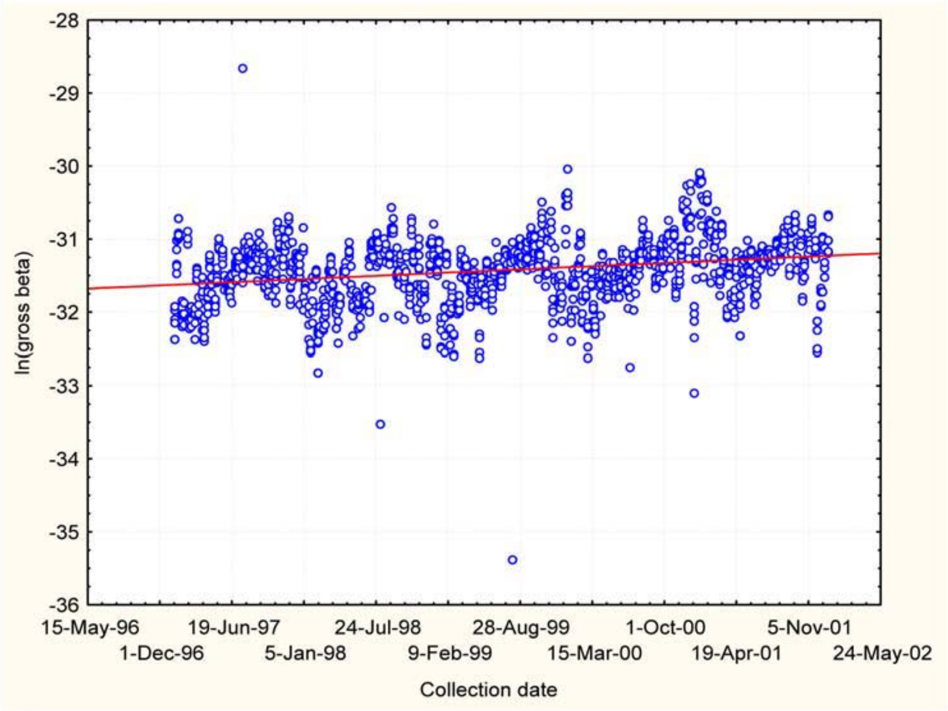

**Figure B-7. Scatter plot and regression line for ln(gross beta) from distant locations.**

**Table B-3. Regression equations and associated statistics for boundary and distant locations.**

| Sample Group | Regression Equation | Correlation Coefficient | p-value |
|---|---|---|---|
| Boundary | ln(gross beta) = -38.7 + 0.245×(date) | 0.245 | <0.0001 |
| Distant | ln(gross beta) = -39.4 + 0.253×(date) | 0.253 | <0.0001 |

## *Comparison of Slopes*

A comparison of slopes between the regression lines for the boundary locations and distant locations will indicate if the rate of change in gross beta over time differs with location. The comparison of slopes can be performed by constructing 95 percent confidence intervals about the slope parameter (Neter, J. and W. Wasserman 1974). If these intervals overlap, we can conclude that there is no evidence to suggest a difference in slopes for the two groups of locations.

A confidence interval for the slope is constructed as

$$b - t_{0.025,n-2} s_b \leq \beta \leq b + t_{0.025,n-2} s_b$$

where

$b$ = point estimate of the slope

$t_{0.025,n-2}$ = the Student's t-value associated with two-sided 95 percent confidence and n-2 degrees of freedom

$s_b$ = the standard deviation of the slope estimate, b

ß = the true slope, which is unknown.

Table B-4 gives the values used in constructing the confidence intervals and the resulting confidence intervals. As seen in the fifth column of Table B-4, the confidence intervals for the slope overlap and we can conclude that there is no difference in the rate of change in gross beta measurements for the two location groupings, boundary and distant.

**Table B-4. Ninety-five percent confidence intervals on the true slope.**

| Sample group | b | z[a] | $s_b$ | 95% C.I.[b] |
|---|---|---|---|---|
| Boundary | 0.245 | 1.96 | 0.0229 | [0.200, 0.290] |
| Distant | 0.253 | 1.96 | 0.0269 | [0.200, 0.306] |

a. For large sample sizes, the standard normal z-value is used instead of the Student's t-value.

b. C.I. = confidence interval.

# REFERENCES

Bartholomay, R.C., B.J. Tucker, L.C. Davis, and M.R. Greene, 2000, Hydrological Conditions and Distribution of Selected Constituents in water, Snake River Plain Aquifer, Idaho National Engineering and Environmental Laboratory, Idaho, 1996 Through 1998, U.S. Geological Survey, DOE/ID-22167. September 2000.

Currie, L.A., 1984, Lower Limit of Detection-Definition and Elaboration of a Proposed Position for Radiological Effluent and Environmental Measurements: U.S. Nuclear Regulatory Commission NUREG/CR-4007.

Hollander, M., and Wolfe, D. A., 1973, *Nonparametric Statistical Methods*, New York: John Wiley and Sons, Inc.

Neter, J. and Wasserman, W., 1974, *Applied Linear Statistical Models*, Homewood, Illinois: Richard D. Irwin, Inc.

Shapiro, S. S., and Wilk, M. B., 1965, "An Analysis of Variance Test for Normality (complete samples)," *Biometrika*, 52, 591-611.

Shapiro, S. S., Wilk, M. B., and Chen, H. J., 1968, "A Comparative Study of Various Tests of Normality," *Journal of the American Statistical Association*, 63, 1343-1372.